Canonical Correlation Analysis

Fall 2014

STAT 560

Dimension reduction for two sets of random variables

When distinction between explanatory and response variables are not so clear, an analysis dealing with two sets of variables in a symmetric manner is desired.

- Relationship between gene expressions and biological variables
- Relation between two sets of psychological tests, each with multidimensional measurements

Job Satisfaction Data

- Dunham investigated the extent to which measures of job satisfaction are related to job characteristics.
- Measurements of $p = 5$ job characteristics and $q = 7$ job satisfaction variables for $n = 784$ executives from the corporate branch of a large retail merchandising corporation were obtained.
- Data sets: $\mathbf{X}_{5\times 784}$, $\mathbf{Y}_{7\times 784}$
- Are measures of job satisfaction associated with job characteristics?

Job Satisfaction Data

- X_1: feedback
- X_2: task significance
- X_3: task variety
- X_4: task identity
- X_5 : autonomy
- Y_1 : supervisor satisfaction
- Y_2 : career-future satisfaction
- Y_3 : financial satisfaction
- Y_4 : workload satisfaction
- Y_5 : company identification
- Y_6 : kind-of-work-satisfaction
- Y_7 : general satisfaction

Job Satisfaction Data

- Interested in dimension reduction of $\mathbf{X}$ and $\mathbf{Y}$, while keeping the important association between $\mathbf{X}$ and $\mathbf{Y}$.
- Find linear dimension reduction of $\mathbf{X}$ and $\mathbf{Y}$ using the cross-correlation matrix.

Correlation Matrix

$$\mathbf{R}_X = \begin{bmatrix} 1 & & & & \\ 0.49 & 1 & & & \\ 0.53 & 0.57 & 1 & & \\ 0.49 & 0.46 & 0.48 & 1 & \\ 0.51 & 0.53 & 0.57 & 0.57 & 1 \end{bmatrix}$$

Correlation Matrix

$$\mathbf{R}_Y = \begin{bmatrix} 1 & & & & & & \\ 0.43 & 1 & & & & & \\ 0.27 & 0.33 & 1 & & & & \\ 0.24 & 0.26 & 0.25 & 1 & & & \\ 0.34 & 0.54 & 0.46 & 0.28 & 1 & & \\ 0.37 & 0.32 & 0.29 & 0.3 & 0.35 & 1 & \\ 0.4 & 0.58 & 0.45 & 0.27 & 0.59 & 0.31 & 1 \end{bmatrix}$$

Correlation Matrix

$$
\mathbf{R}_{X,Y} = \left[\begin{array}{ccccc|ccccccc}
1 & & & & & & & & & & & \\
0.49 & 1 & & & & & & & & & & \\
0.53 & 0.57 & 1 & & & & & & & & & \\
0.49 & 0.46 & 0.48 & 1 & & & & & & & & \\
0.51 & 0.53 & 0.57 & 0.57 & 1 & & & & & & & \\
\hline
0.33 & 0.3 & 0.31 & 0.24 & 0.38 & 1 & & & & & & \\
0.32 & 0.21 & 0.23 & 0.22 & 0.32 & 0.43 & 1 & & & & & \\
0.2 & 0.16 & 0.14 & 0.12 & 0.17 & 0.27 & 0.33 & 1 & & & & \\
0.19 & 0.08 & 0.07 & 0.19 & 0.23 & 0.24 & 0.26 & 0.25 & 1 & & & \\
0.3 & 0.27 & 0.24 & 0.21 & 0.32 & 0.34 & 0.54 & 0.46 & 0.28 & 1 & & \\
0.37 & 0.35 & 0.37 & 0.29 & 0.36 & 0.37 & 0.32 & 0.29 & 0.3 & 0.35 & 1 & \\
0.21 & 0.2 & 0.18 & 0.16 & 0.27 & 0.4 & 0.58 & 0.45 & 0.27 & 0.59 & 0.31 & 1
\end{array}\right]
$$

CCA

- CCA seeks to identify and quantify the associations between two sets of variables.
- Given two random vectors $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$, consider linear combination of each of two random vectors,

$$U = \mathbf{a}^\top \mathbf{X} = a_1 X_1 + \cdots + a_p X_p$$
$$V = \mathbf{b}^\top \mathbf{Y} = b_1 Y_1 + \cdots + b_q Y_q.$$

- CCA finds the random variables (U, V) or the direction vectors $(\mathbf{a}, \mathbf{b})$ which provide maximal correlation between U and V,

$$Corr(U, V) = \frac{Cov(U, V)}{\sqrt{Var(U) Var(V)}}.$$

Population CCA

- Denote the full covariance between $(\mathbf{X}, \mathbf{Y})$ by

$$\begin{aligned}\Sigma_{X,Y} &= \left[\begin{array}{c|c} Cov(\mathbf{X}) & Cov(\mathbf{X}, \mathbf{Y}) \\ \hline Cov(\mathbf{Y}, \mathbf{X}) & Cov(\mathbf{Y}) \end{array}\right] \\ &= \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right]\end{aligned}$$

- The first set of direction vectors solve

$$\begin{aligned}(\mathbf{a}_1, \mathbf{b}_1) &= \arg\max_{\mathbf{a}\in\mathbb{R}^p, \mathbf{b}\in\mathbb{R}^q} Corr(\mathbf{a}^\top \mathbf{X}, \mathbf{y}^\top \mathbf{Y}) \\ &= \arg\max_{\mathbf{a}\in\mathbb{R}^p, \mathbf{b}\in\mathbb{R}^q} \frac{\mathbf{a}^\top \Sigma_{12} \mathbf{b}}{\sqrt{\mathbf{a}^\top \Sigma_{11} \mathbf{a} \mathbf{b}^\top \Sigma_{22} \mathbf{b}}}\end{aligned}$$

- $(\mathbf{a}_1, \mathbf{b}_1)$: Canonical correlation vectors
- $(U_1 = \mathbf{a}_1^\top \mathbf{X}, V_1 = \mathbf{b}_1^\top \mathbf{Y})$: canonical variables
- $\rho_1 = Corr(U_1, V_1)$: canonical correlation

Population CCA

- The subsequent canonical correlation vectors given $(\mathbf{a}_1, \ldots, \mathbf{a}_{k-1})$ and $(\mathbf{b}_1, \ldots, \mathbf{b}_{k-1})$ are

$$(\mathbf{a}_k, \mathbf{b}_k) = \arg\max_{\substack{\mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^q \\ \mathbf{a}^\top \Sigma_{11} \mathbf{a}_j = 0 \\ \mathbf{b}^\top \Sigma_{22} \mathbf{b}_j = 0 \\ j = 1, \ldots, k-1}} Corr(\mathbf{a}^\top \mathbf{X}, \mathbf{y}^\top \mathbf{Y})$$

- $(\mathbf{a}_k, \mathbf{b}_k)$: k-th Canonical correlation vectors
- $(U_k = \mathbf{a}_k^\top \mathbf{X}, V_k = \mathbf{b}_k^\top \mathbf{Y})$: k-th canonical variables
- $\rho_k = Corr(U_k, V_k)$: k-th largest canonical correlation
- $Corr(U_k, U_j) = 0$, $Corr(V_k, V_j) = 0$, for $j = 1, \ldots, k-1$.
- In general, $\mathbf{a}_k^\top \mathbf{a}_j = 0$ $\mathbf{b}_k^\top \mathbf{b}_j = 0$ NOT true.

Sample CCA

- Replace population variance-covariance and cross-covariance matrices by the sample matrices.
- Denote the sample covariance matrices by $\mathbf{S}_{ij} = \hat{\Sigma}_{ij}$ for $i, j = 1, 2$.
- The, the sample CCA is

$$(\mathbf{a}_k, \mathbf{b}_k) \quad = \quad \arg\max_{\substack{\mathbf{a} \in \mathbb{R}^p,\, \mathbf{b} \in \mathbb{R}^q \\ \mathbf{a}^\top \mathbf{S}_{11} \mathbf{a}_j = 0 \\ \mathbf{b}^\top \mathbf{S}_{22} \mathbf{b}_j = 0 \\ j = 1, \ldots, k-1}} \frac{\mathbf{a}^\top \mathbf{S}_{12} \mathbf{b}}{\sqrt{\mathbf{a}^\top \mathbf{S}_{11} \mathbf{a} \mathbf{b}^\top \mathbf{S}_{22} \mathbf{b}}}$$

How to solve CCA problem?

- Change of variable: $\mathbf{c} = \mathbf{S}_{11}^{1/2}\mathbf{a}, \mathbf{d} = \mathbf{S}_{22}^{1/2}\mathbf{b}$
- Then, we are maximizing

$$\frac{\mathbf{c}^\top \mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{2}^{-1/2}\mathbf{d}}{\sqrt{\mathbf{c}^\top\mathbf{c}\mathbf{d}^\top\mathbf{d}}}$$

- One can verify that the solution $(\mathbf{c}_k, \mathbf{d}_k)$ is given by the eigen-directions

$$\mathbf{C} = \mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2}$$
$$\mathbf{D} = \mathbf{S}_{22}^{-1/2}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2}.$$

- Also, $(\lambda_k, \mathbf{c}_k)$ is eigenvalue-vector pair for $\mathbf{C}$, then $(\lambda_k, \mathbf{d}_k)$ is for $\mathbf{D}$.
- Then, $\mathbf{a} = \mathbf{S}_{11}^{-1/2}\mathbf{c}$ and $\mathbf{b} = \mathbf{S}_{22}^{-1/2}\mathbf{d}$ and $\rho_1^2 = \lambda_1$.

Finding CCA solution

- Find the eigenvectors $\mathbf{c}_1, \ldots, \mathbf{c}_d$ and $\mathbf{d}_1, \ldots, \mathbf{d}_d$ of

$$\mathbf{C} = \mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2}$$
$$\mathbf{D} = \mathbf{S}_{22}^{-1/2}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2},$$

  where $d = \min(p, q)$.
- The k-th CCA direction vectors then given by

$$\mathbf{a}_k = \mathbf{S}_{11}^{-1/2}\mathbf{c}_k, \mathbf{b}_k = \mathbf{S}_{22}^{-1/2}\mathbf{d}_k$$

  and $\rho_k^2 = \lambda_k$.

Finding CCA solution

- Or equivalently, find the eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_d$ and $\mathbf{b}_1, \ldots, \mathbf{b}_d$ of

$$\mathbf{A} = \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$$
$$\mathbf{B} = \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12},$$

  where $d = \min(p, q)$.
- Then, $(\mathbf{a}_k, \mathbf{b}_k)$ are the k-th CCA direction vectors and $\rho_k^2 = \lambda_k$.

Job satisfaction data- CCA

- Eigenvalues of $\mathbf{A} = \mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$:

$$0.3066,\ 0.0559,\ 0.0142,\ 0.0052,\ 0.0033$$

- Eigenvectors of $\mathbf{A} = \mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$:

$$\begin{bmatrix} -0.6246 & -0.2455 & -0.5865 & -0.5536 & -0.0205 \\ -0.2890 & 0.4786 & 0.3032 & -0.1890 & -0.6524 \\ -0.2483 & 0.6110 & -0.1773 & 0.3291 & 0.6066 \\ 0.0339 & -0.2550 & -0.2893 & 0.7318 & -0.3479 \\ -0.6808 & -0.5219 & 0.6701 & -0.1181 & 0.2915 \end{bmatrix}$$

Job satisfaction data- CCA

- Eigenvalues of $\mathbf{B} = \mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$:

$$0.3066,\ 0.0559,\ 0.0142,\ 0.0052,\ 0.0033$$

- Eigenvectors of $\mathbf{B} = \mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$:

$$\begin{bmatrix} 0.5542 & -0.0748 & -0.3319 & 0.1081 & -0.3889 & 0.4641 & -0.6318 \\ 0.2723 & 0.3710 & 0.5285 & 0.2867 & -0.6046 & -0.3330 & 0.3915 \\ -0.0468 & -0.0790 & 0.3225 & 0.5102 & 0.2788 & 0.5910 & -0.4346 \\ 0.0307 & 0.7875 & 0.0044 & -0.3405 & 0.2512 & 0.1547 & -0.1887 \\ 0.3784 & -0.0860 & -0.1910 & 0.3762 & 0.5668 & -0.5431 & 0.2472 \\ 0.6722 & -0.4714 & 0.2784 & -0.5789 & 0.1448 & -0.0703 & 0.1427 \\ -0.1436 & -0.0270 & -0.6265 & -0.2306 & -0.0117 & -0.0250 & 0.3764 \end{bmatrix}$$

First Canonical variate pair

The first sample canonical variate pair is

$$\hat{U}_1 = -0.62X_1 - 0.29X_2 - 0.25X_3 + 0.03X_4 - 0.68X_5$$

and

$$\hat{V}_1 = 0.55Y_1 + 0.27Y_2 - 0.05Y_3 + 0.03Y_4 + 0.38Y_5 + 0.67Y_6 - 0.14Y_7$$

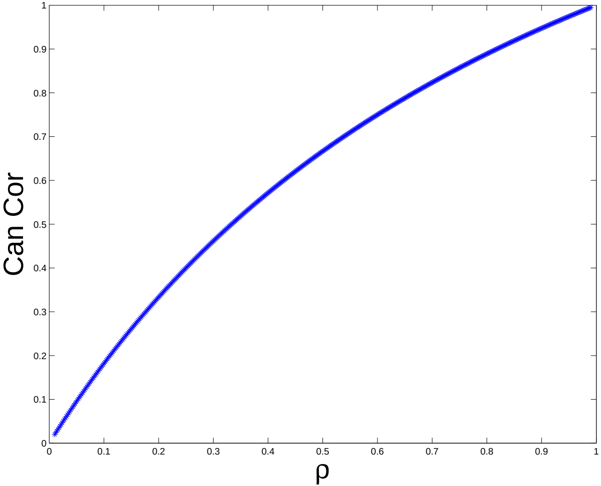
with the sample canonical correlation, $\hat{\rho}_1 = \sqrt{\lambda_1} = 0.55$.

(Population) Canonical Correlation Example

Determine the first canonical variate pair for the bivariate variables $X = (X_1, X_2)^T$ and $Y = (Y_1, Y_2)^T$ whose correlation matrix is given as follows:

$$\mathbf{R} = \left[\begin{array}{cc|cc} 1 & & & \\ \rho & 1 & & \\ \hline \rho & \rho & 1 & \\ \rho & \rho & \rho & 1 \end{array}\right]$$

(Population) Canonical Correlation Example



Large Sample Inference

- When $\Sigma_{12} = 0$, $Corr(a^T X, b^T Y)$ have covariance $a^T \Sigma_{12} b = 0$ for any vectors a and b. All canonical correlations must be zero, and there is no point in pursuing a CCA.
- LRT of
$$H_0 : \Sigma_{12} = 0, \quad H_a : \Sigma_{12} \neq 0$$
rejects H_0 for large values of
$-2\log(\text{likelihood ratio}) = -n\, log \prod_{i=1}^{r}(1 - \hat{\rho}_i^2)$.
- Bartlett suggested replacing the multiplicative factor n in the likelihood ratio statistic with the factor $n - 1 - \frac{1}{2}(p + q + 1)$ to improve the χ^2 approximation to the sampling distribution of $-2\log\Lambda$.
- Thus, for large n, reject H_0 at significance level α if
$$-2\log\tilde{\Lambda} = -\left(n - 1 - \frac{1}{2}(p+q+1)\right) log \prod_{i=1}^{r}(1 - \hat{\rho}_i^2) > \chi^2_{pq}(\alpha).$$

Large Sample Inference: Job satisfaction data

- For the job satisfaction data, we have $-2\log\tilde{\Lambda} = 350$ and $\chi^2_{pq}(0.05) = 49.8$, thus, reject the null concluding that the data strongly supports non-zero canonical correlations.
- The canonical correlations are ordered from the largest to the smallest, we can conclude that the first canonical correlation is nonzero.
- What about the subsequent canonical correlations?

Large Sample Inference

- Consider

$$H_0^k : \rho_1 \neq 0, \cdots \rho_k \neq 0, \rho_{k+1} = \cdots = \rho_r = 0$$

vs

$$H_0^k : \rho_i \neq 0 \text{ for some } i \geq k+1.$$

- For large n, reject H_0^k if

$$-\left(n - 1 - \frac{1}{2}(p+q+1)\right) log \prod_{i=k+1}^{r} (1 - \hat{\rho}_i^2) > \chi^2_{(p-k)(q-k)}(\alpha).$$

Large Sample Inference: Job satisfaction data

- Consider
$$H_0^1 : \rho_1 \neq 0, \rho_2 = \cdots = \rho_r = 0$$
vs
$$H_0^1 : \rho_i \neq 0 \text{ for some } i \geq 2.$$
- Reject H_0^1 since
$$\begin{aligned} -\left(n - 1 - \frac{1}{2}(p + q + 1)\right) \log \prod_{i=2}^{r} (1 - \hat{\rho}_i^2) &= 62.98 \\ > \chi^2_{(p-1)(q-1)}(0.05) &= 36.42. \end{aligned}$$

Canonical Correlation: Qualitative Data

	Subject's status				
Father's status	1	2	3	4	5
1	50	45	8	18	8
2	28	174	84	154	55
3	11	78	110	223	96
4	14	150	185	714	447
5	0	42	72	320	411

Table : Example: Social mobility contingency table

Canonical Correlation: Social mobility Data

- $X = (X_1, \ldots, X_5)$: X_i: dummy variable indicating subject being in the i-th social status
- $Y = (Y_1, \ldots, Y_5)$: Y_i: dummy variable indicating father being in the i-th social status
- Look for $\mathbf{a}, \mathbf{b} \in \mathbf{R}^5$ which maximizes the correlation between $\mathbf{a}^T X$ and $\mathbf{b}^T Y$.
- Let $\mathbf{X}$ and $\mathbf{Y}$ be the dummy data matrices corresponding to the contingency table, one for row and other for column categories.
- Define the grand data matrix $\mathbf{W} = [\mathbf{X}, \mathbf{Y}]$.

Canonical Correlation: Social mobility Data

- Note that

$$n\mathbf{S_W} = \begin{pmatrix} \mathrm{diag}(\mathbf{f}) - \mathbf{f}\mathbf{f}^T/n & \mathbf{N} - \hat{\mathbf{N}} \\ \mathbf{N}^T - \hat{\mathbf{N}}^T & \mathrm{diag}(\mathbf{g}) - \mathbf{g}\mathbf{g}^T/n \end{pmatrix},$$

  where
  - $\mathbf{N}$: the contingency table
  - $\mathbf{f} = (\sum_i n_{ij})$: column vector for marginal row sum
  - $\mathbf{g} = (\sum_j n_{ij})$: column vector for marginal column sum
  - $\hat{\mathbf{N}} = \mathbf{f}\mathbf{g}^T/n$: expected contingency table under the the assumption that the row and column categories are independent.
- The submatrices $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ are not invertible.

Canonical Correlation: Social mobility Data

- Let $\mathbf{S}_{ij}^*$ be the matrix by deleting the first columns/row vector of $\mathbf{S}_{ij}$.
- Similarly, $\mathbf{f}^*$ and $\mathbf{g}^*$ be the vectors obtained by deleting the first components of $\mathbf{f}$ and $\mathbf{g}$.
- Then, one can show that $(n\mathbf{S}_{11}^*)^{-1} = [\text{diag}(\mathbf{f}^*)]^{-1} + f_1^{-1}\mathbf{1}\mathbf{1}^T$, $(n\mathbf{S}_{22}^*)^{-1} = [\text{diag}(\mathbf{g}^*)]^{-1} + g_1^{-1}\mathbf{1}\mathbf{1}^T$.
- The CCA directions can be obtained from the eigenvectors of

$$\mathbf{A} = (\mathbf{S}_{11}^*)^{-1}\mathbf{S}_{12}^*(\mathbf{S}_{22}^*)^{-1}\mathbf{S}_{21}^*$$

and

$$\mathbf{B} = (\mathbf{S}_{22}^*)^{-1}\mathbf{S}_{21}^*(\mathbf{S}_{11}^*)^{-1}\mathbf{S}_{12}^*.$$

Canonical Correlation: Qualitative Data

- The first eigenvectors of **A** and **B**

$$\mathbf{a}_1 = [0.37, 0.49, 0.54, 0.58]$$

$$\mathbf{b}_1 = [0.37, 0.49, 0.54, 0.58]$$

corresponding to the largest eigenvalue: 0.2537.

- The first sample canonical variate pair is:

$$\hat{U}_1 = 0X_1 + 0.37X_2 + 0.49X_3 + 0.54X_4 + 0.58X_5$$

and

$$\hat{V}_1 = 0Y_1 + 0.37Y_2 + 0.49Y_3 + 0.54Y_4 + 0.58Y_5$$

with the first sample canonical correlation $\hat{\rho}_1 = \sqrt{.2537} = 0.5037$.

- Both father's and son's social class, the coefficients appear in their natural order.
- Social classes 1 and 2 seem to be more distinct from on another than the other adjacent social classes, both for the son's and the father's status.